NOTE ON A FORMULA FOR THE MULTIPLE CORRELATION COEFFICIENT

By H. M. Bacon

There are many useful formulas available for the calculation of the multiple correlation coefficient in a k variable problem.[1] Since it frequently happens that the regression equation is the primary object of the statistical analysis, the well known formula

$$r^2_{1\cdot 23\cdots k} = \beta_{12\cdot 34\cdots k}r_{12} + \beta_{13\cdot 24\cdots k}r_{13} + \cdots + \beta_{1k\cdot 23\cdots (k-1)}r_{1k}$$

can be used to considerable advantage. While many different demonstrations of this formula are perfectly familiar, the one given in this note may prove of some interest.

First let us recapitulate briefly certain facts about the regression coefficients and the multiple correlation coefficient. Suppose we have k sets of N numbers each:

$$\begin{array}{ccccc} X_{11} & X_{12} & \cdot & \cdot & X_{1N} \\ X_{21} & X_{22} & \cdot & \cdot & X_{2N} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ X_{k1} & X_{k2} & \cdot & \cdot & X_{kN}\,. \end{array}$$

Let $\bar{x}_j$ be the mean of the j-th set, and let $x_{ji} = X_{ji} - \bar{x}_j$. We then have k sets of N deviations from means, and we shall suppose the following k sets to be linearly independent:

$$\begin{array}{ccccc} x_{11} & x_{12} & \cdot & \cdot & x_{1N} \\ x_{21} & x_{22} & \cdot & \cdot & x_{2N} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{k1} & x_{k2} & \cdot & \cdot & x_{kN}\,. \end{array}$$

We shall consider only the regression of the "variable" x_1 upon $x_2, x_3, \cdots, x_k$. Clearly the results obtained can be made to describe the regression of any one of the variables upon the other $k - 1$ variables by rearranging the subscripts. As usual let $\lambda_2, \lambda_3, \cdots, \lambda_k$ have values which will make the sum of squares

$$F(\lambda_2, \lambda_3, \cdots, \lambda_k) = \Sigma(x_{1i} - \lambda_2 x_{2i} - \lambda_3 x_{3i} - \cdots - \lambda_k x_{ki})^2$$

a minimum. For simplicity we shall omit stating limits of summation and understand hereafter that Σ means "sum for i from $i = 1$ to $i = N$." Neces-

[1] For example, see W. J. Kirkham, "Note on the Derivation of the Multiple Correlation Coefficient", *The Annals of Mathematical Statistics*, Volume VIII (1937), pp. 68–71.

sary conditions (which are easily shown to be sufficient) are that λ_2, λ_3, $\cdots$; λ_k must satisfy the equations

$$(1)\qquad \begin{aligned} \frac{\partial F}{\partial \lambda_2} &= -2\Sigma(x_{1i} - \lambda_2 x_{2i} - \lambda_3 x_{3i} - \cdots - \lambda_k x_{ki})x_{2i} = 0 \\ \frac{\partial F}{\partial \lambda_3} &= -2\Sigma(x_{1i} - \lambda_2 x_{2i} - \lambda_3 x_{3i} - \cdots - \lambda_k x_{ki})x_{3i} = 0 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \frac{\partial F}{\partial \lambda_k} &= -2\Sigma(x_{1i} - \lambda_2 x_{2i} - \lambda_3 x_{3i} - \cdots - \lambda_k x_{ki})x_{ki} = 0. \end{aligned}$$

These equations are simply the "normal equations" for determining the regression coefficients. Solving them we obtain

$$\begin{aligned} \lambda_2 &= b_{12\cdot 34\cdots k} \\ \lambda_3 &= b_{13\cdot 24\cdots k} \\ &\cdot\;\cdot\;\cdot\;\cdot\;\cdot\;\cdot \\ \lambda_k &= b_{1k\cdot 23\cdots (k-1)}\,. \end{aligned}$$

The equation of regression of x_1 on x_2, x_3, $\cdots$, x_k is therefore

$$x_1 = b_{12\cdot 34\cdots k}x_2 + b_{13\cdot 24\cdots k}x_3 + \cdots + b_{1k\cdot 23\cdots (k-1)}x_k\,.$$

If we let

$$u_i = b_{12\cdot 34\cdots k}x_{2i} + b_{13\cdot 24\cdots k}x_{3i} + \cdots + b_{1k\cdot 23\cdots (k-1)}x_k$$

for $i = 1, 2, \cdots, N$ then $x_{1i} - u_i$ is the residual of the i-th x_1. The coefficient of multiple correlation of x_1 in terms of x_2, x_3, $\cdots$, x_k is defined to be the simple correlation coefficient of the x's and u's:

$$r_{1\cdot 234\cdots k} = \frac{\Sigma x_{1i}u_i}{\sqrt{\Sigma x_{1i}^2 \Sigma u_i^2}}.$$

In case it is desired to express the x's in terms of their standard deviations, the following equation is used:

$$\frac{x_1}{\sigma_1} = \beta_{12\cdot 34\cdots k}\frac{x_2}{\sigma_2} + \beta_{13\cdot 24\cdots k}\frac{x_3}{\sigma_3} + \cdots + \beta_{1k\cdot 23\cdots (k-1)}\frac{x_k}{\sigma_k}$$

or

$$z_1 = \beta_{12\cdot 34\cdots k}z_2 + \beta_{13\cdot 24\cdots k}z_3 + \cdots + \beta_{1k\cdot 23\cdots (k-1)}z_k$$

where

$$(2)\qquad \begin{aligned} \beta_{12\cdot 34\cdots k} &= b_{12\cdot 34\cdots k}\frac{\sigma_2}{\sigma_1} \\ \beta_{13\cdot 24\cdots k} &= b_{13\cdot 24\cdots k}\frac{\sigma_3}{\sigma_1} \\ &\cdots\cdots\cdots\cdots\cdots \\ \beta_{1k\cdot 23\cdots (k-1)} &= b_{1k\cdot 23\cdots (k-1)}\frac{\sigma_k}{\sigma_1} \end{aligned}$$

and

$$z_j = \frac{x_j}{\sigma_j}.$$

Now if $\Sigma A_i B_i = 0$, the set of numbers $A_1, A_2, \cdots, A_N$ is said to be *orthogonal* to the set of numbers $B_1, B_2, \cdots, B_N$. Hence the conditions of equations (1) may be described by saying that the values of $\lambda_2, \lambda_3, \cdots, \lambda_k$ must be such that the set of residuals $x_{1i} - u_i$ is orthogonal to each of the $k - 1$ sets of numbers $x_{2i}, x_{3i}, \cdots, x_{ki}$. But if the set of residuals is orthogonal to each of these sets, it is orthogonal to any linear combination of them. Since the set of u's is such a linear combination, we have

$$\Sigma(x_{1i} - u_i)u_i = 0$$

and hence

$$\Sigma x_{1i} u_i = \Sigma u_i^2 . \tag{3}$$

Since $u_i = b_{12\cdot34\cdots k}x_{2i} + b_{13\cdot24\cdots k}x_{3i} + \cdots + b_{1k\cdot23\cdots(k-1)}x_{ki}$ it follows at once by multiplying both sides by x_{1i} and summing that

$$\Sigma x_{1i}u_i = b_{12\cdot34\cdots k}\Sigma x_{1i}x_{2i} + b_{13\cdot24\cdots k}\Sigma x_{1i}x_{3i} + \cdots + b_{1k\cdot23\cdots(k-1)}\Sigma x_{1i}x_{ki} . \tag{4}$$

Writing

$$\begin{aligned} \Sigma x_{1i}x_{2i} &= N\sigma_1\sigma_2 r_{12} \\ \Sigma x_{1i}x_{3i} &= N\sigma_1\sigma_3 r_{13} \\ &\cdots\cdots\cdots \\ \Sigma x_{1i}x_{ki} &= N\sigma_1\sigma_k r_{1k} , \end{aligned}$$

noting the relations between the b's and the β's expressed in equations (2), and observing that we may write

$$\Sigma x_{1i}u_i = \frac{(\Sigma x_{1i}u_i)^2}{\Sigma x_{1i}u_i} = \frac{(\Sigma x_{1i}u_i)^2}{\Sigma u_i^2}$$

because of equation (3), we may therefore rewrite equation (4) as follows

$$\frac{(\Sigma x_{1i}u_i)^2}{\Sigma u_i^2} = \beta_{12\cdot34\cdots k}\frac{\sigma_1}{\sigma_2}N\sigma_1\sigma_2 r_{12} + \beta_{13\cdot24\cdots k}\frac{\sigma_1}{\sigma_3}N\sigma_1\sigma_3 r_{13} + \cdots$$
$$\cdots + \beta_{1k\cdot23\cdots(k-1)}\frac{\sigma_1}{\sigma_k}N\sigma_1\sigma_k r_{1k}.$$

Now divide both sides by $\Sigma x_{1i}^2 = N\sigma_1^2$ obtaining

$$r_{1\cdot234\cdots k}^2 = \frac{(\Sigma x_{1i}u_i)^2}{\Sigma x_{1i}^2 \Sigma u_i^2} = \beta_{12\cdot34\cdots k}r_{12} + \beta_{13\cdot24\cdots k}r_{13} + \cdots + \beta_{1k\cdot23\cdots(k-1)}r_{1k}.$$

This is the formula which was to be established.

Stanford University.